

Recent advances in bibliometric indexes and the *PaperRank* problem[★]

Pierluigi Amodio^{a,*} and Luigi Brugnano^b

^a*Dipartimento di Matematica, Università di Bari, Italy*

^b*Dipartimento di Matematica “U. Dini”, Università di Firenze, Italy*

Abstract

Bibliometric indexes are customary used in evaluating the impact of scientific research, even though it is very well known that in different research areas they may range in very different intervals. Sometimes, this is evident even within a single given field of investigation making very difficult (and inaccurate) the assessment of scientific papers. On the other hand, the problem can be recast in the same framework which has allowed to efficiently cope with the ordering of web-pages, i.e., to formulate the *PageRank* of Google. For this reason, we call such problem the *PaperRank* problem, here solved by using a similar approach to that employed by *PageRank*. The obtained solution, which is mathematically grounded, will be used to compare the usual heuristics of the number of citations with a new one here proposed. Some numerical tests show that the new heuristics is much more reliable than the currently used ones, based on the bare number of citations. Moreover, we show that our model improves on recently proposed ones [3].

Key words: Bibliometric indexes, *PageRank*, citations, *H-index*, normalized citations.

1 Introduction

In recent years, it has become the fashion to evaluate the impact of research by using bibliometric indexes. This approach clearly does not solve the problem

[★] Work developed within the project “*Innovative Methods of Numerical Linear Algebra with Applications*”.

* Corresponding author.

Email addresses: amodio@dm.uniba.it (Pierluigi Amodio),
luigi.brugnano@unifi.it (Luigi Brugnano).

at hand, though it can be useful to have a gross idea about specific issues. For instance, the quality of a scientist is sometimes ranked by using the so called *H-index* [11], though it is very well known that this can be useful only for analysing short time return researches, whereas it could be completely inadequate to assess more basic research fields which, in turn, prove to be important (and, sometimes, priceless) only after decades or centuries. As an example, the *little Fermat theorem*, on which modern electronic secure transactions essentially rely, dates back to 1640, when it was apparently useless. Also, very important mathematicians are known to have very small *H-index* (e.g., Galois has an *H-index* equal to 2 ...).

Nevertheless, in some circumstances, bibliometric indexes allow to obtain a rough idea about the impact of research, though their value heavily depends on the chosen index. In particular, it is recognized that the bare number of citations is a parameter which has many drawbacks: it does depend on the specific field of research, on unfair behaviors (which, unfortunately, are not unknown in the scientific setting), etc.

On the other hand, this problem is known to be structurally similar to that of ranking urls on the web. As is well known, this latter problem has been formalized in the *Google PageRank* [5] (see also [16]). Based on the simple idea that the importance of a web page depends on the number of web pages that link to it and on their relative importance, the *PageRank* relies on a solid mathematical basis which allows the search engine Google [10] to efficiently recover information across the web (see also [17,7] for a deeper mathematical analysis of the corresponding matrix problem, and [2,3] for generalizations). Approaches based on this idea have been used for evaluating the impact of scientific journals (see, e.g., [1,4]) and the impact of scientific articles (see, e.g., [6,15,14,13]), also taking into account of collaborations [8]. Sometimes, however, the above procedures are not mathematically well refined. In any case, such approaches use a *global information* which could be difficult to recover and manage efficiently (it is enough thinking to the computation of the *Google PageRank* to realize the possible complexity of the problem). Indeed, numerical algorithms need to be finely tuned, in order to gain efficiency (see, e.g., [9,18]). This is the main reason why heuristics, like the number of citations (which are relatively easy to compute), have become popular, even though, as pointed out above, sometimes they may provide misleading advices. Consequently, more efficient heuristics would be desirable for dealing with the problem.

With this premise, in this paper we provide the model for constructing a mathematically grounded ranking of what we call the *PaperRank* problem¹ which, under some mild assumptions, is proved to exist and to be unique. The results

¹ In analogy with the *PageRank* problem of the web.

of this model are more reliable than those given by the model recently proposed in [3], and will be therefore assumed as “reference solutions” to validate a new heuristics, of local nature. The numerical examples here provided then clearly show that in many significant instances the new heuristics is much more reliable and fair than the usual one based on the bare number of citations.

2 The *Random Reader* Model for the *PaperRank* Problem

The principle that we here describe is the analogous of that used in [5] for deriving the famous *PageRank* of Google. For this reason, we name the problem *PaperRank* problem. We would like to remind that the mathematically based theory underlying the definition of the *Google PageRank* is the reason for its effectiveness in retrieving informations across the web. In the present setting, its basic principle may be then reformulated as follows:

“an important paper is cited by important papers.”

That is, in analogy with the *random surfer* model proposed for the *PageRank* problem, we now have a virtual *random reader*, which starts reading a paper, then randomly passing to read a paper cited in it. If we repeat this process indefinitely, the *importance* of a given paper is the fraction of time that the random reader spends in reading it (assuming, obviously, that each paper is read in a constant time). This principle can be formally modeled by introducing the following *citation matrix*²

$$L = (\ell_{ij}) \in \mathbb{R}^{N \times N}, \quad \ell_{ij} = \begin{cases} 1 & \text{if paper } j \text{ cites paper } i \\ 0 & \text{otherwise} \end{cases}, \quad \forall i, j = 1, \dots, N. \quad (1)$$

Remark 1 *We observe that, by introducing the unit vector*

$$\mathbf{e} = (1, \dots, 1)^T \in \mathbb{R}^N,$$

then the vector containing the number of citations of each paper is given by

$$\mathbf{c} = \begin{pmatrix} c_1 \\ \vdots \\ c_N \end{pmatrix} \equiv L \mathbf{e} \quad (2)$$

² Actually, by looking at the papers as the nodes of an oriented graph, such a matrix is nothing but the transpose of the adjacency matrix.

(i.e., c_i is the number of citations of the i th paper). Such vector is currently used for computing several bibliometric indexes such as, e.g., the H -index.

On the other hand, the vector

$$\mathbf{f}^T = (f_1, \dots, f_N) \equiv \mathbf{e}^T L$$

contains the number of bibliographic items in each paper. That is, f_j is the number of references in paper j , $\forall j = 1, \dots, N$. If we then define v_i as the *importance* of the i th paper, then

$$v_i = \sum_{j=1}^N \ell_{ij} v_j f_j^+, \quad i = 1, \dots, N, \quad f_j^+ = \begin{cases} f_j^{-1} & \text{if } f_j > 0 \\ 0 & \text{otherwise} \end{cases}.$$

By introducing the vectors

$$\mathbf{v} = (v_1, \dots, v_N)^T, \quad \mathbf{f}^+ = (f_1^+, \dots, f_N^+)^T,$$

and the matrices

$$F = \text{diag}(\mathbf{f}), \quad F^+ = \text{diag}(\mathbf{f}^+), \quad (3)$$

the previous set of equations can be cast in vector form as

$$\mathbf{v} = L F^+ \mathbf{v} \equiv S \mathbf{v}. \quad (4)$$

However, this ranking could not exist or might be not unique, depending whether $1 \in \sigma(S)$, and/or if this eigenvalue is simple.

In order to cope with this problem, in [3] the authors introduce, in a similar model, a *dummy paper*, say 0, which references all the other ones and is referenced by all of them.³ That is, matrix L is replaced by the augmented matrix

$$\hat{L} = \begin{pmatrix} 0 & \mathbf{e}^T \\ \mathbf{e} & L \end{pmatrix} \in \mathbb{R}^{N+1 \times N+1}. \quad (5)$$

Matrix \hat{S} as in (4) is then defined accordingly:

$$\hat{S} = \hat{L} \hat{F}^+ \equiv \hat{L} \begin{pmatrix} N^{-1} & \\ & (I + F)^{-1} \end{pmatrix}, \quad (6)$$

with F the diagonal matrix defined in (3) and I the identity matrix of dimension N . Moreover, matrix (6) is clearly *irreducible*.

³ We here consider only the problem of ranking the papers, whereas in [3] a more general problem is modeled.

However, this last feature, makes the model not very faithful, in that it is quite well known that there exist groups of papers, whose citations do not overlap, so that matrix L is indeed *reducible*. This is often the case, for example, of different fields of research within the same discipline or in different ones. For this reason, we here propose a different solution to this problem, in which we assume that, by default, *each paper references itself*, that is $\ell_{ii} = 1$, for all $i = 1, \dots, N$, so that (see (3))

$$f_j \geq 1, \quad j = 1, \dots, N, \quad \implies \quad F^+ = F^{-1}.$$

Consequently,

$$\mathbf{e}^T S = \mathbf{e}^T L F^+ = \mathbf{f}^T F^{-1} = \mathbf{e}^T,$$

so that $1 \in \sigma(S)$ and, evidently, the possible reducibility of the original matrix (1) is retained by the modified one.

Concerning the fact that ranking is unique, by following similar steps as those used for the *Google PageRank*, we may assume that, having reached a given paper, the *random reader* chooses with probability $p \in (0, 1)$ a paper cited in it, or it *jumps* to read at random *any* paper, with probability $1 - p$. In vector form this reads:

$$\mathbf{v} = S(p)\mathbf{v} \equiv \left(pS + \frac{1-p}{N} \mathbf{e}\mathbf{e}^T \right) \mathbf{v}, \quad p \in (0, 1). \quad (7)$$

Since

$$S(p) > 0, \quad \|S(p)\|_1 = 1, \quad \forall p \in (0, 1),$$

from the Perron-Frobenius Theorem (see, e.g., [12]), one easily deduces that $1 \in \sigma(S(p))$, which is a simple eigenvalue, separating in modulus all other eigenvalues of $S(p)$. In addition, the corresponding eigenvector $\mathbf{v} > 0$. We then conclude that the *PaperRank* problem (7) admits a solution, which is feasible (i.e., with positive entries) and unique. Moreover, by choosing $p \approx 1$, $S(p) \approx S(1) \equiv S$ and, therefore, the approximate model well matches the original one (see also Section 2.1 below).

Consequently, this ranking is *rigorously mathematically grounded*, though it requires an information of global nature, alike the case for computing the *Google PageRank*. This means that it is relatively costly, since it requires to know all the data about every bibliographical item.

Remark 2 *We observe that this last step (i.e., the introduction of the parameter p) is not required for the matrix \hat{S} (see (6)) of the model derived from (5), since one easily proves the following result.*

Theorem 1 *Let $L \neq 0$ and \hat{S} defined according to (6). Then $\hat{S}^4 > 0$.*

In other words, there always exists a path of exact length 4 between any two of the nodes $0, \dots, N$, provided that $L \neq 0$.

2.1 Perturbation analysis

In this section we provide a simple analysis showing how the introduction of the parameter p in (7) affects the original vector. For this purpose, let us denote the eigenvector as $\mathbf{v}(p)$. That is,

$$S(p)\mathbf{v}(p) = \mathbf{v}(p), \quad p \in (0, 1).$$

Clearly, $\mathbf{v}^* \equiv \mathbf{v}(1)$ is the correct limit vector, which obviously exists, whereas $\mathbf{v}(0) = \frac{1}{N}\mathbf{e}$. Consequently, an estimate for $\mathbf{v}'(p)$ is given by

$$\mathbf{v}' \approx \mathbf{v}(1) - \mathbf{v}(0) = \mathbf{v}^* - \frac{1}{N}\mathbf{e}.$$

One then obtains that, for $p \approx 1$:

$$\mathbf{v}(p) \approx \mathbf{v}(1) + (p - 1)\mathbf{v}' = p\mathbf{v}^* + \frac{1 - p}{N}\mathbf{e}. \quad (8)$$

From (8) one then concludes that the introduction of the parameter p results in an almost uniform (small) perturbation of the entries of the correct vector. As a matter of fact, the statistical properties of the two vectors are practically the same, for all the test problems reported in Section 3.

2.2 A New Heuristics for the PaperRank Problem

As it has been shown in the previous section, the correct *PaperRank* is obtained by starting from the scaled matrix LF^+ , in place of L . Similarly, instead of considering the bare number of citations, given by the vector (2), we propose to use *normalized* citations, defined as the entries of the vector (see (4))

$$\mathbf{c}_{norm} = LF^+\mathbf{e} \equiv S\mathbf{e}, \quad (9)$$

which requires, as (2), only information of local nature. In other words, in place of counting the number of citation to a given paper, we propose to consider the *number of citations to that paper, each divided by the number of references in the corresponding paper containing the citation itself*. It is obvious that the index (9) has the same complexity as (2). Nevertheless, in the next section we show that the statistical properties of (9) are more fair than those of (2), in the sense that they better reproduce the correct ones provided by the reference model (7).

It is evident, from the definition (9), that the vector \mathbf{c}_{norm} is essentially the first iterate of the power method applied to S , by starting from a constant vector. Consequently,

- it requires only a *local* information, even though it would aim to approximate, in the limit, a global one (i.e., the *PaperRank*);
- no more than one iteration is possible, without requiring a global information.

Consequently, the heuristics (9) is the best we can do by using only local information. Nonetheless, as is shown in the numerical tests, it proves to be quite effective.

Remark 3 *It is worth mentioning that the use of the normalized citations (9) also copes correctly with the problem of self-citations. Indeed, for each new published paper, the normalized additional (self-)citations of an author cannot exceed 1. On the contrary, they are virtually unbounded for the vector (2) of bare citations.*

3 Numerical Tests

We here provide a few numerical tests, each one modeling a significant situation, to compare the *PaperRank* obtained from (7) by choosing $p = 0.99$ (which we assume to be the reference one), with those given by the model (2), based on bare citations, and (9), based on the normalized citations. For each test we plot three histograms which rank the papers according to the vectors representing these indexes. For ease of a direct comparison, the obtained vectors are normalized so that their values range in the interval $[0, 1]$.

Moreover, for each problem we also compare the *PaperRank* obtained from (7) with that obtained from (6), that is from the model proposed in [3]. In fact, we shall see that they may significantly differ, due to the fact that our model preserves the possible reducibility of the matrix (1). On the other hand, it is clear that the vectors (2) and (9) derived from the two matrices (4) and (6) are essentially the same (obviously, by neglecting the first entry of the vector, in the second case).

Examples 1 and 2

We suppose to have a single and homogeneous group of 500 articles (first example) or two distinct and homogeneous groups with 300 and 700 articles, respectively (second example). In both cases each paper has a mean of 20 randomly distributed references in its own group (see the first two plots in Figures 1 and 2). In both the examples, all the three rankings (7), (2), and (9) have a similar distribution of the relevance of the papers, as is shown in the

last three plots in Figures 1 and 2, even though (9) better fits the distribution of (7). In Figure 2 the green bars concern the first group of papers, whereas the blue ones concern the second group.

For these problems, the *PaperRanks* obtained from (7) and (6) turn out to be similar each other, so that we do not report the latter ones. Indeed, in the first example, the matrix is irreducible and in the second example reducibility is not an important feature, since the two blocks have similar properties (i.e., the same number of mean references in each paper).

Examples 3 and 4

In these examples, we have two distinct and homogeneous groups of papers, with 300 and 700 items, respectively. Each paper only cites articles in its own group. In the Example 3, each paper in the first group has a mean of 10 randomly distributed references, whereas each paper in the second group has a mean of 70 randomly distributed references (see the first two plots in Figure 3). In Example 4, the situation is reversed since the papers in the smaller group have a mean of 70 randomly distributed citations whereas each paper in the second group has a mean of 10 randomly distributed references (see the first two plots in Figure 4). In the last three plots of Figures 3 and 4, the green bars concern the first group of papers, whereas the blue ones concern the second group. The rankings (7) and (9) always have a similar distribution of the relevance of the papers, whereas the ranking (2) exhibits two peaks (one for the first group and one for the second group), which exchange in the two cases. It is clear that in these situations the different number of references in the papers of each group invalidate the ranking (2), whereas it doesn't affect the normalized ranking (9).

As one may expect, for these problems, there is a significant difference between the *PaperRanks* obtained from our model (7) and that derived from (6). Indeed, in both cases reducibility turns out to be an important feature of the corresponding citation matrices. This is shown in Figures 5 and 6, respectively, for the two examples. In each figure, the upper plot reproduces the third subplot in Figures 3 and 4, respectively, whereas the lower plot depicts the corresponding *PaperRank* derived from (6). It is evident that the latter one does not allow to properly compare the papers in the two groups.

Example 5

In this example, we have two groups of papers: a larger one, with 900 papers, which reference randomly a mean of 20 papers in the same group, and a

smaller one of 100 papers, which reference papers in the same group, with a mean of 50 citations, and the papers in the larger one, with a mean of 20 citations (see the first two plots in Figure 7). In this case, the heuristics (2), based on the bare number of citations, recognizes two groups of papers, the most important being the smaller one (the green bars in the fourth plot of Figure 7). The correct distribution, however, is that depicted in the third plot of Figure 7, given by (7), where the green bars are the leftmost (i.e., the less important) ones. This behaviour is qualitatively better reproduced in the last plot of Figure 7, concerning the heuristics (9).

For this problem, the *PaperRanks* obtained from (7) and (6) turn out to be similar each other, since the citation matrix turns out to be irreducible. Consequently, we do not report the latter one.

Example 6

The last example concerns the case of three groups of papers:

- a group of 200 *leader* papers, which randomly reference a mean of 20 papers in the same group;
- a group of 200 papers, which randomly reference a mean of 20 *leader* papers and 20 papers in its own group;
- a group of 400 papers, which randomly reference a mean of 20 *leader* papers and 100 papers in its own group;

This situation is summarized by the first two plots in Figure 8. It is evident that the correct ranking is that depicted in the third plot of Figure 8, representing the vector in (7), with the *leader* papers (in red) more important than those in the second group (in green) and those in the third group (in blue), these latter having the same importance. This situation is qualitatively well reproduced by the new heuristics (9), as is shown in the last plot of Figure 8, where the *leader* papers (red) are again the most important, and the other ones (green and blue) have a comparable importance, though the blue ones are slightly oversized. Vice versa, the usual ranking (2), based on the bare number of citations (which is shown in the fourth plot of Figure 8), depicts a wrong scenario, in which the *leader* papers are replaced by those with the highest number of internal references (third group).

For this problem, the *PaperRanks* obtained from (7) and (6) turn out to be similar each other, since the citation matrix turns out to be irreducible. Consequently, we do not report the latter one.

4 Conclusions

In this paper, we provide a mathematically correct definition of the *Paper-Rank* problem to assess scientific papers, which is able to properly compare also papers in disjoint groups, thus improving on a recent model proposed in [3]. On the basis of this new model, we provide a local heuristics, based on normalized citations, which appears to be quite effective (though much cheaper to compute), allowing to overcome some well known drawbacks of the ranking based on the bare number of citations.

References

- [1] C.T. Bergstrom. Eigenfactor: Measuring the value and prestige of scholarly journals. *College & Research Libraries News* **68**, 5 (2007) 314–316.
- [2] D.A. Bini, G.M. Del Corso, F. Romani. Evaluating scientific products by means of citation-based models: a first analysis and validation. *Electron. Trans. Numer. Anal.* **33** (2008/09) 1–16.
- [3] D.A. Bini, G.M. Del Corso, F. Romani. A combined approach for evaluating papers, authors and scientific journals. *Jour. of Comput. and Appl. Math.* **234** (2010) 3104–3121.
- [4] J. Bollen, H. Van de Sompel, A. Hagberg, R. Chute. A Principal Component Analysis of 39 Scientific Impact Measures. *PLoS ONE* **4** 6 (2009) e6022.
- [5] S. Brin, L. Page. The Anatomy of a Large-Scale Hypertextual Web Search Engine. (1998) <http://dbpubs.stanford.edu/pub/1998-8>
- [6] P. Chen, H. Xie, S. Maslov, S. Redner. Finding scientific gems with google’s pagerank algorithm. *Journal of Informetrics* **1** (2007) 8–15.
- [7] A. Cicone, S. Serra-Capizzano. Google PageRanking problem: The model and the analysis. *Journal of Computational and Applied Mathematics* **234** (2010) 3140–3169.
- [8] Y. Ding, E. Yan, A. Frazho, J. Caverlee. PageRank for Ranking Authors in Co-citation Networks. *Jour. of the American Soc. for Information Sci. and Technology* **60** 11 (2009) 2229–2243.
- [9] G.H. Golub. An Arnoldi-type algorithm for computing the Page Rank. *BIT* **46** (2006) 759–771.
- [10] <http://www.google.it/>
- [11] J.E. Hirsch. An index to quantify an individuals scientific research output. *PNAS* **102** 46 (2005) 16569–16572.

- [12] P. Lancaster, M. Tismenetsky. *The theory of matrices. Second edition.* Academic Press, 1985.
- [13] J. Li, P. Willett. ArticleRank: a PageRank-based alternative to numbers of citations for analysing citation networks. *Aslib Proceedings* **61** 6 (2009) 605–618.
- [14] N. Ma, J. Guan, Y. Zhao. Bringing PageRank to the citation analysis. *Information Processing and Management* **44** (2008) 800–810.
- [15] S. Maslov, S. Redner. 2008. Promise and Pitfalls of Extending Google PageRank Algorithm to Citation Networks. *The Journal of Neuroscience* **28** 44 (2008) 11103–11105.
- [16] L. Page, S. Brin, R. Motwani, T. Winograd. *The PageRank Citation Ranking: Bringing Order to the Web.* Technical Report, Stanford InfoLab, 1999 <http://ilpubs.stanford.edu:8090/422/>
- [17] S. Serra-Capizzano. Jordan canonical form of the Google matrix: a potential contribution to the PageRank computation. *SIAM J. Matrix Anal. Appl.* **27**, 2 (2005) 305–312.
- [18] J.-F. Yin, G.-J. Yin, M. Ng. On adaptively accelerated Arnoldi method for computing PageRank. *Numer. Linear Algebra Appl.* **19** (2012) 73–85.

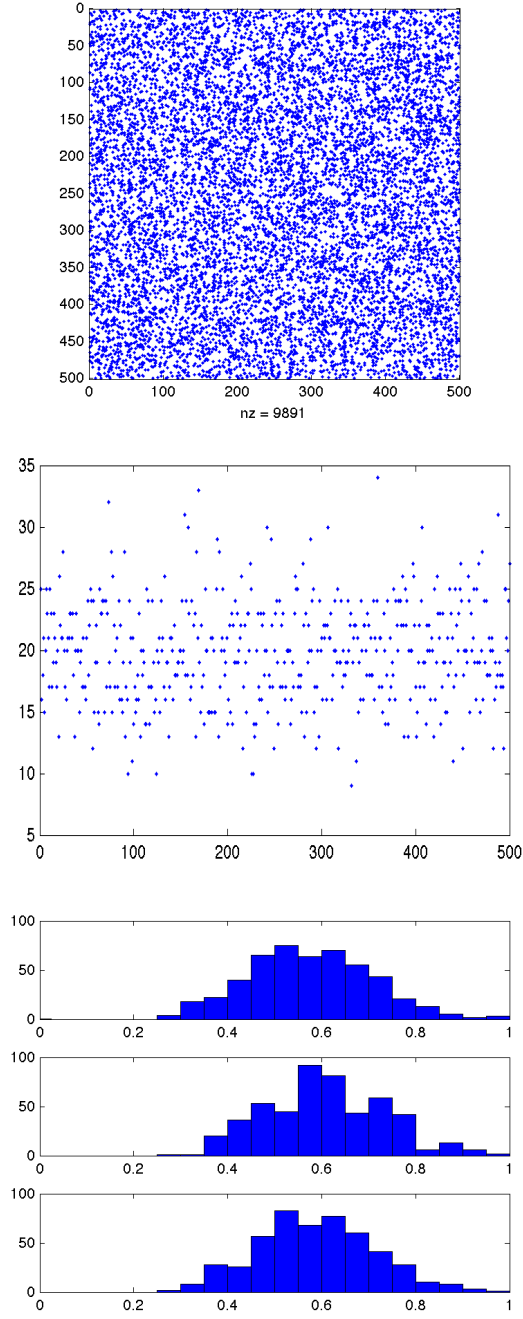


Fig. 1. citation matrix (1), number of references (2), and rankings (7), (2), and (9) for Example 1, respectively.

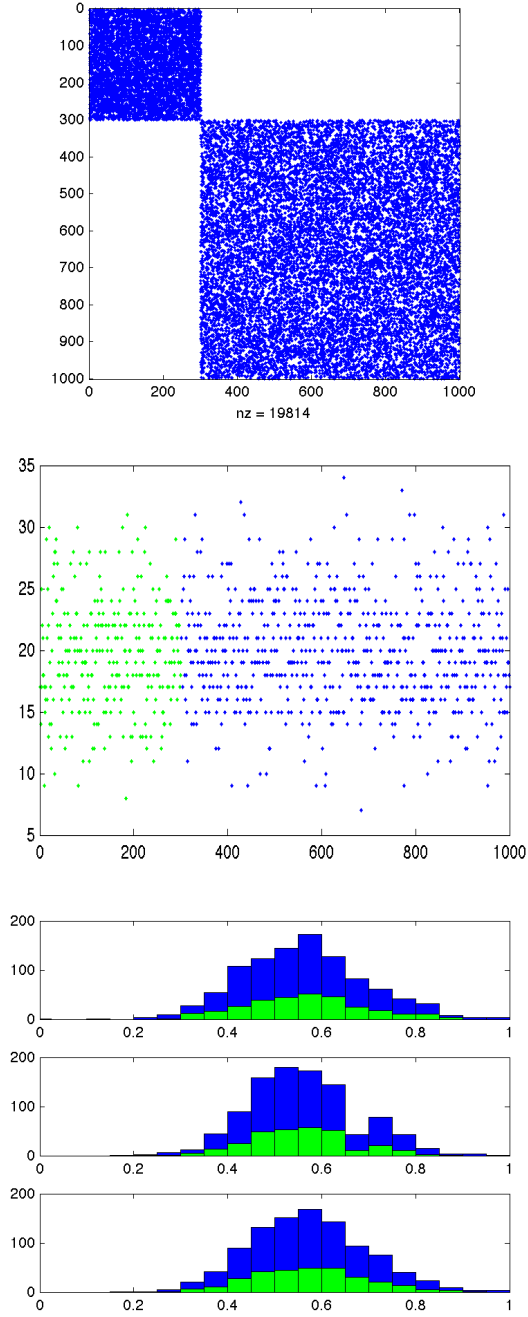


Fig. 2. citation matrix (1), number of references (2), and rankings (7), (2), and (9) for Example 2, respectively.

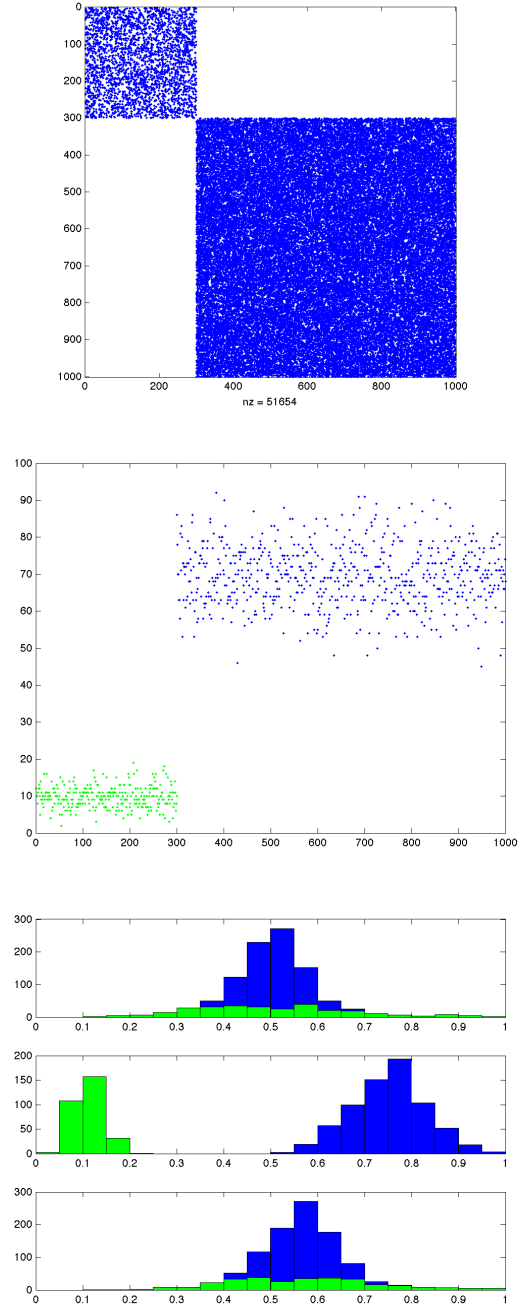


Fig. 3. citation matrix (1), number of references (2), and rankings (7), (2), and (9) for Example 3, respectively.

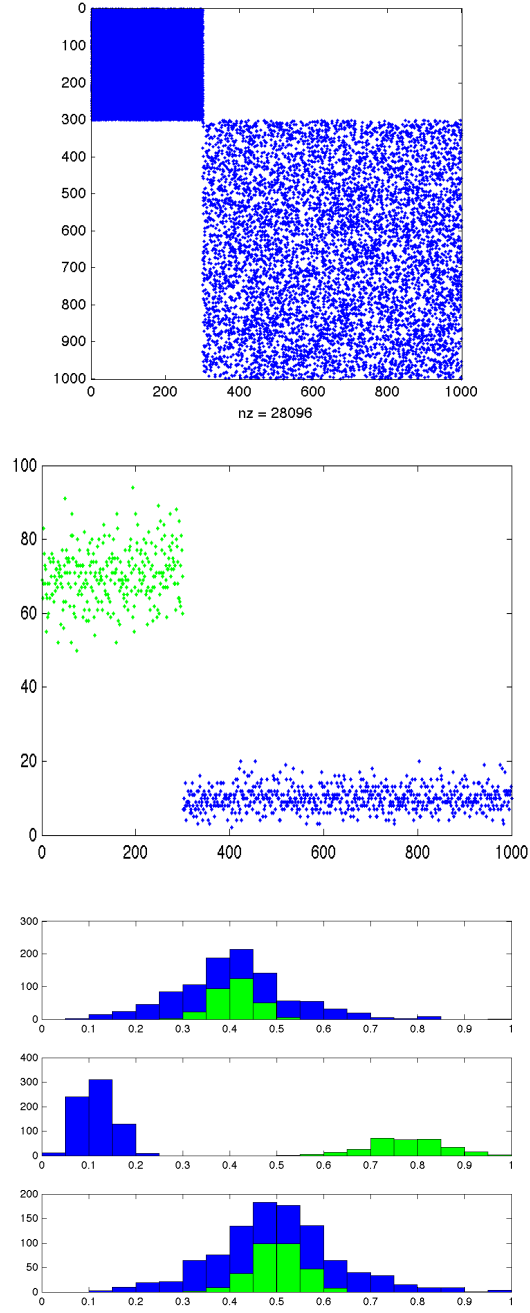


Fig. 4. citation matrix (1), number of references (2), and rankings (7), (2), and (9) for Example 4, respectively.

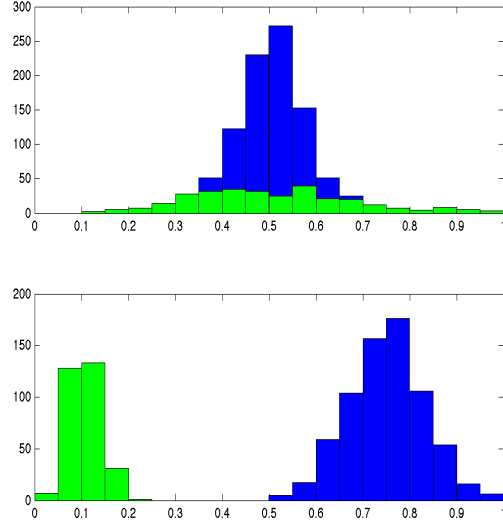


Fig. 5. *PaperRanks* derived from (7) (upper plot) and (6) (lower plot) for Example 3.

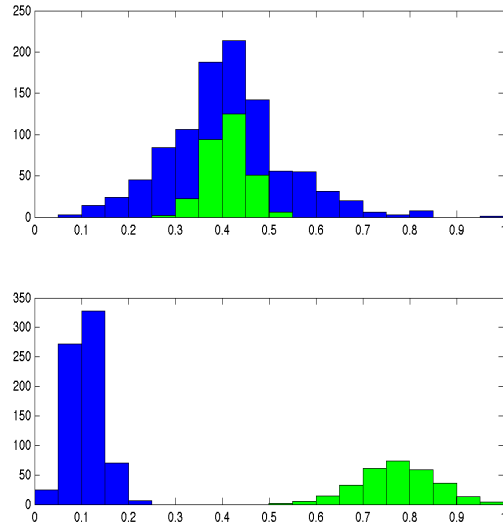


Fig. 6. *PaperRanks* derived from (7) (upper plot) and (6) (lower plot) for Example 4.

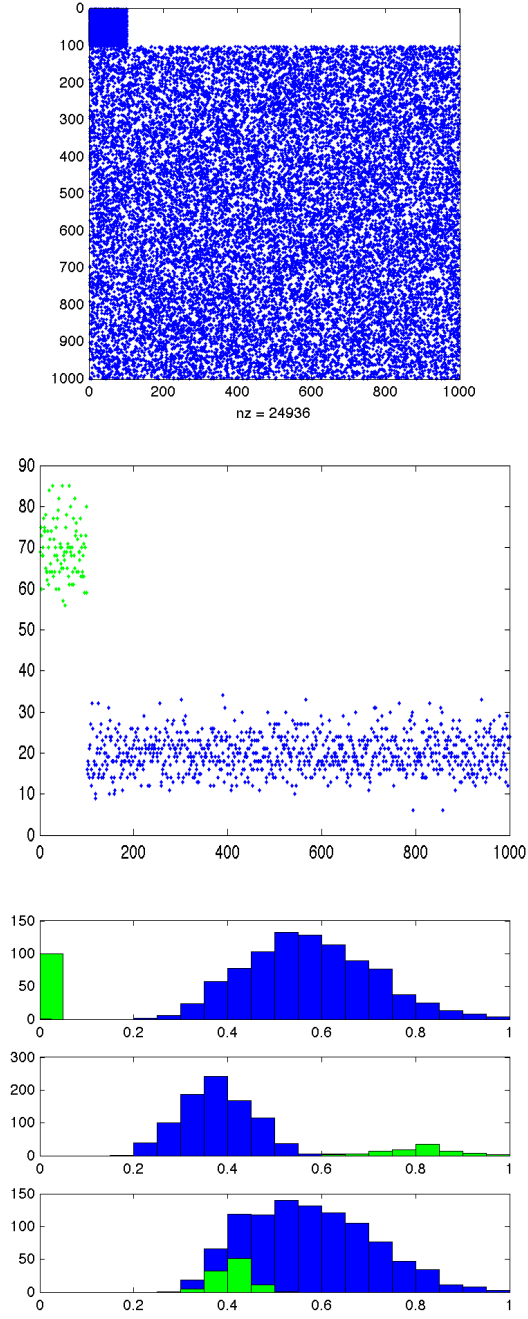


Fig. 7. citation matrix (1), number of references (2), and rankings (7), (2), and (9) for Example 5, respectively.

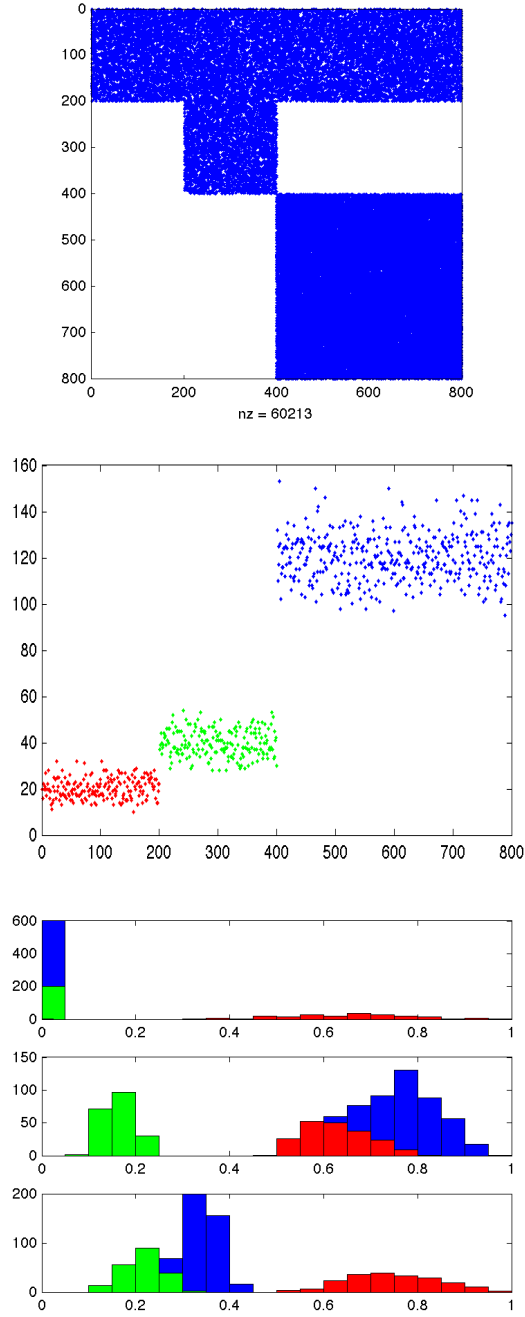


Fig. 8. citation matrix (1), number of references (2), and rankings (7), (2), and (9) for Example 6, respectively.